第二章

条件期望和线性投影

Outline

1. 条件期望
2. 条件方差
3. 线性投影

条件期望的定义

令 $\boldsymbol{X}$ 为 K 维随机向量，Y 为一元随机变量。以 $\boldsymbol{X}$ 为条件，Y 的条件期望函数（Conditional Expectation Function, CEF）定义如下：

$$\mu(\boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X}) = \begin{cases} \int_Y y \cdot f(y \mid \boldsymbol{X}) dy, & Y \text{ 连续}, \\ \sum_Y y \cdot \mathbb{P}_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}), & Y \text{ 离散}, \end{cases}$$

其中，$f(y \mid \boldsymbol{X})$ 是以 $\boldsymbol{X}$ 为条件 $Y = y$ 的条件概率密度函数，$\mathbb{P}_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})$ 是以 $\boldsymbol{X}$ 为条件 $Y = y$ 的条件概率。

条件期望的性质

简单迭代期望律（Simple Law of Iterated Expectations, SLIE）

若 $\mathbb{E}|Y| < \infty$，则对任意的随机向量 $\boldsymbol{X}$，我们有：

$$\mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[Y \mid \boldsymbol{X}]] = \mathbb{E}[Y],$$

其中，符号 $\mathbb{E}_{\boldsymbol{X}}[\cdot]$ 表示相对于随机向量 $\boldsymbol{X}$ 求期望。

证明.

假设 Y 和 $\boldsymbol{X}$ 的联合概率密度函数为 $f(y,\boldsymbol{x})$，因为 $\mathbb{E}[Y\mid\boldsymbol{X}]$ 是 $\boldsymbol{X}$ 的函数，我们有：

$$\mathbb{E}[\mathbb{E}[Y\mid\boldsymbol{X}]]=\int_{\mathbb{R}^K}\mathbb{E}[Y\mid\boldsymbol{X}=\boldsymbol{x}]\cdot f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}.$$

将 Y 为连续随机变量情况下条件期望的定义代入上式得到：

$$\begin{aligned}&\int_{\mathbb{R}^K}\left(\int_{\mathbb{R}}yf(y\mid\boldsymbol{x})dy\right)f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}=\int_{\mathbb{R}^k}\int_{\mathbb{R}}yf(y,\boldsymbol{x})dyd\boldsymbol{x}\\=&\int_{\mathbb{R}}yf_Y(y)dy=\mathbb{E}[Y],\end{aligned}$$

其中，第一个等号成立是因为 $f(y\mid\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})=f(y,\boldsymbol{x})$；第二个等号成立是因为联合分布和边缘分布之间存在如下关系：
$\int_{\mathbb{R}^K}f(y,\boldsymbol{x})d\boldsymbol{x}=f_Y(y)$。 □

迭代期望律

若 $\mathbb{E}|Y| < \infty$，则对任意的随机向量 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$，我们有：

$$\mathbb{E}_{\boldsymbol{X}_1}[\mathbb{E}[Y \mid \boldsymbol{X}_1, \boldsymbol{X}_2] \mid \boldsymbol{X}_1] = \mathbb{E}[Y \mid \boldsymbol{X}_1].$$

实际上，可以证明如下更一般形式的迭代期望律：令 $\boldsymbol{W}$ 为随机向量，$\boldsymbol{X}$ 是随机向量且满足 $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{W})$。[1] 一般形式的 LIE 可表示为：

$$\mathbb{E}[Y \mid \boldsymbol{X}] = \mathbb{E}\left[\mathbb{E}(Y \mid \boldsymbol{W}) \mid \boldsymbol{X}\right].$$

"The smaller information set always dominates."

[1] 如果已知 $\boldsymbol{W}$ 的实现值，$\boldsymbol{X}$ 的实现值便是已知的。

CEF 分解性质

对任意的随机变量 Y，有如下性质成立：

$$Y = \mathbb{E}[Y \mid \boldsymbol{X}] + \varepsilon, \tag{1}$$

其中，

(a) ε 均值独立于 $\boldsymbol{X}$，即 $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$；

(b) ε 与 $\boldsymbol{X}$ 的任意函数都不相关。

证明.

首先证明 (a),

$$\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}\left\{Y - \mathbb{E}[Y \mid \boldsymbol{X}] \mid \boldsymbol{X}\right\} = \mathbb{E}[Y \mid \boldsymbol{X}] - \mathbb{E}[Y \mid \boldsymbol{X}] = 0.$$

其次证明 (b),假设 $h(\boldsymbol{X})$ 是 $\boldsymbol{X}$ 的任意函数,我们有:

$$\begin{aligned}
\operatorname{Cov}[\varepsilon, h(\boldsymbol{X})] &= \mathbb{E}[\varepsilon \cdot h(\boldsymbol{X})] - \mathbb{E}(\varepsilon)\mathbb{E}[h(\boldsymbol{X})] \\
&= \mathbb{E}\{\mathbb{E}[\varepsilon \cdot h(\boldsymbol{X}) \mid \boldsymbol{X}]\} - 0 \cdot \mathbb{E}[h(\boldsymbol{X})] \\
&= \mathbb{E}\{h(\boldsymbol{X}) \cdot \mathbb{E}[\varepsilon \mid \boldsymbol{X}]\} \\
&= 0.
\end{aligned}$$

□

CEF 预测性质

假设 $\mathbb{E}(Y^2) < \infty$，$\boldsymbol{X}$ 为 K 维随机向量，记 $\mu(\boldsymbol{X}) \equiv \mathbb{E}(Y \mid \boldsymbol{X})$，则 μ 是如下优化问题的解：

$$\min_{m(\boldsymbol{X})} \mathbb{E}[(Y - m(\boldsymbol{X}))^2],$$

其中，$m : \mathbb{R}^K \to \mathbb{R}$ 满足 $\mathbb{E}[m(\boldsymbol{X})^2] < \infty$。我们称 $\mathbb{E}[(Y - m(\boldsymbol{X}))^2]$ 为 $m(\boldsymbol{X})$ 预测 Y 时的均方误差（mean squared error, MSE）。

条件方差

令 $\boldsymbol{X}$ 为随机向量，Y 为一元随机变量。如果 $\mathbb{E}[Y^2] < \infty$，以 $\boldsymbol{X} = \boldsymbol{x}$ 为条件，Y 的条件方差（Conditional Variance）定义如下：

$$\sigma^2(\boldsymbol{x}) = \mathrm{Var}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}])^2 \mid \boldsymbol{X} = \boldsymbol{x}\right].$$

条件方差作为随机变量 $\boldsymbol{X}$ 的函数可记为 $\sigma^2(\boldsymbol{X})$。

方差分解

若 $\mathbb{E}[Y^2] < \infty$，则有：

$$\mathrm{Var}[Y] = \mathrm{Var}\left[\mathbb{E}\left[Y \mid \boldsymbol{X}\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[Y \mid \boldsymbol{X}\right]\right].$$

证明.

根据 CEF 分解性质，我们有：

$$\mathrm{Var}[Y] = \mathrm{Var}\left\{\mathbb{E}\left[Y \mid \boldsymbol{X}\right] + \varepsilon\right\}.$$

根据 CEF 分解性质中的 (b)，

$$\mathrm{Var}\left\{\mathbb{E}\left[Y \mid \boldsymbol{X}\right] + \varepsilon\right\} = \mathrm{Var}\left[\mathbb{E}\left(Y \mid \boldsymbol{X}\right)\right] + \mathrm{Var}(\varepsilon).$$

现只需证明：$\mathrm{Var}(\varepsilon) = \mathbb{E}\left[\mathrm{Var}\left(Y \mid \boldsymbol{X}\right)\right]$。事实上，根据 CEF 分解性质中的 (a) 和 SLIE，$\mathbb{E}(\varepsilon) = 0$。于是，我们有：

$$\begin{aligned}\mathrm{Var}(\varepsilon) = \mathbb{E}(\varepsilon^2) &= \mathbb{E}\left[\mathbb{E}\left(\varepsilon^2 \mid \boldsymbol{X}\right)\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[\left(Y - \mathbb{E}(Y \mid \boldsymbol{X})\right)^2 \mid \boldsymbol{X}\right]\right\} \\ &= \mathbb{E}\left[\mathrm{Var}\left(Y \mid \boldsymbol{X}\right)\right],\end{aligned}$$

其中，第二个等号成立用到了 SLIE；第三个等号成立用到了 $\varepsilon = Y - \mathbb{E}(Y \mid \boldsymbol{X})$；最后一个等号成立用到了条件方差的定义。 □

线性投影

- 虽然条件期望 $\mu(\boldsymbol{X})$ 提供了 Y 的最优预测，但是 $\mu(\boldsymbol{X})$ 的函数形式通常是未知的。
- 在实证分析时，一般假设 CEF 具有线性形式，但是这种假设往往不切实际。更准确的做法是将线性形式的函数设定看作是关于 CEF 的近似。

线性投影

令 $\boldsymbol{X}$ 为随机向量。以 $\boldsymbol{X}$ 为条件，关于 Y 的最优线性预测，也称为线性投影，被定义为：

$$\mathscr{P}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta},$$

其中，$\boldsymbol{\beta}$ 被称为线性投影系数，也称为总体回归系数，它最小化如下均方预测误差：

$$S(\boldsymbol{b}) = \mathbb{E}\left((Y - \boldsymbol{X}'\boldsymbol{b})^2\right).$$

$S(\boldsymbol{b})$ 的最小化问题也被称为总体最小二乘（population least squares）问题。

线性投影的性质

(1) 线性投影系数 β 存在且唯一，其表达式为:

$$\boldsymbol{\beta} = (\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'])^{-1}\mathbb{E}[\boldsymbol{X}Y].$$

(2) 投影误差 $e = Y - \boldsymbol{X}'\boldsymbol{\beta}$ 满足 $\mathbb{E}[e^2] < \infty$ 和 $\mathbb{E}[\boldsymbol{X}e] = \mathbf{0}$。特别地，若 $\boldsymbol{X}$ 中含有常数项，则 $\mathbb{E}(e) = 0$。$\mathbb{E}[\boldsymbol{X}e] = \mathbf{0}$ 被称为 $\boldsymbol{X}$ 与 e 正交。

线性投影与 CEF

(1) 假设 CEF 是 $\boldsymbol{X}$ 的线性函数，则 $\mu(\boldsymbol{X})=\mathscr{P}(Y\mid\boldsymbol{X})$.

(2) $\mathscr{P}(Y\mid\boldsymbol{X})$ 是关于 Y 的最小化 MSE 的最优线性预测。

(3) $\mathscr{P}(Y\mid\boldsymbol{X})$ 是 CEF 的最优（最小化 MSE）线性近似，即 β 是如下最优化问题的解：

$$\min_{\boldsymbol{b}}\mathbb{E}\left\{\left(\mathbb{E}\left[Y\mid\boldsymbol{X}\right]-\boldsymbol{X}'\boldsymbol{b}\right)^2\right\}.$$

考虑两种特殊情况：

- $\boldsymbol{X} = (1, X_1)'$，在此种情况下，我们有：

$$Y = \alpha + \beta_1 X_1 + e,$$

其中，α 和 β_1 分别是常数项 1 和 X_1 对应的线性投影系数。容易证明：

$$\beta_1 = \frac{\mathrm{Cov}(Y, X_1)}{\mathrm{Var}(X_1)}, \quad \alpha = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X_1). \tag{2}$$

- $\boldsymbol{X} = (1, X_1, X_2, \ldots, X_K)'$，$K > 1$，在此种情况下，我们有：

$$\beta_k = \frac{\mathrm{Cov}(Y, \tilde{X}_k)}{\mathrm{Var}(\tilde{X}_k)}, \tag{3}$$

其中，β_k 是 X_k 对应的线性投影系数，$\tilde{X}_k$ 是 X_k 对 $\boldsymbol{X}$ 中所有其他变量投影（包括常数项 1）得到的投影误差。

首先，将 Y 写成线性投影加投影误差的形式：

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_K X_K + e.$$

将上式代入 $\mathrm{Cov}(Y, \tilde{X}_k)$，我们有：

- $\mathrm{Cov}(\tilde{X}_k, e) = 0$，因为 $\tilde{X}_k$ 是 $\boldsymbol{X}$ 各个分量的线性组合，e 与 $\boldsymbol{X}$ 不相关；
- $\mathrm{Cov}(\tilde{X}_k, X_j) = 0$, $j \neq k$，因为 $\tilde{X}_k$ 是线性投影误差，与 X_j 不相关；
- $\mathrm{Cov}(\tilde{X}_k, X_k) = \mathrm{Cov}(\tilde{X}_k, \tilde{X}_k + \boldsymbol{X}'_{-k}\boldsymbol{\beta}_{-k}) = \mathrm{Var}(\tilde{X}_k)$，其中，$\boldsymbol{X}_{-k}$ 表示除 X_k 之外的所有协变量构成的向量，即 $\boldsymbol{X}_{-k} = (1, X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K)'$，$\boldsymbol{\beta}_{-k}$ 是 X_k 对 $\boldsymbol{X}_{-k}$ 进行投影得到的线性投影系数。

于是，

$$\frac{\mathrm{Cov}(Y, \tilde{X}_k)}{\mathrm{Var}(\tilde{X}_k)} = \frac{\beta_k \mathrm{Var}(\tilde{X}_k)}{\mathrm{Var}(\tilde{X}_k)} = \beta_k.$$

References

Joshua D. Angrist & Jörn-Steffen Pischke (2009)
Mostly Harmless Econometrics: An Empiricsist's Companion, Chapter 3.
Princeton University Press, 2009.

Bruce E. Hansen. (2020)
Econometrics, Section 2.18.
https://www.ssc.wisc.edu/ bhansen/econometrics/ Link